

DOCUMENT RESUME

ED 235 504

CS 207 900

AUTHOR Zorn, Jeffrey L.
TITLE Possible Sources of Culture Bias in the Validation of ETS Language Tests.
PUB DATE Mar 83
NOTE 10p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (34th, Detroit, MI, March 17-19, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Viewpoints (120)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; Admission Criteria; Aptitude Tests; College Admission; College Applicants; *Cultural Differences; Higher Education; Objective Tests; *Predictive Validity; Standardized Tests; *Test Bias; Testing; *Testing Problems; Test Reviews; Test Validity; *Writing Evaluation; Writing Skills
IDENTIFIERS English Composition Test; Scholastic Aptitude Test; Test of Standard Written English

ABSTRACT

Although by traditional measures of test validity, the verbal section of the Scholastic Aptitude Test (SAT) is not culture biased, it, along with the English Composition Test and the Test of Standard Written English, reinforces a narrow view of academic excellence that excludes culturally different youths. Designed to prevent admission errors, the SAT is generally successful at helping colleges eliminate students likely to fail in their programs. According to a recent study, however, the test also encourages schools to reject students who would have succeeded. While the SAT does accurately predict freshman grades, it assumes that freshman grades themselves constitute a significant and unbiased criterion of achievement. The objective English Composition Test and the Test of Standard Written English both assume that grammatical correctness is a more important indicator of writing ability than the quality of ideas presented. If colleges are not to exclude a large number of potentially successful students, they must reexamine the suppositions underlying their admissions and placement tests. (MM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Jeffrey L. Zorn

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Possible Sources of Culture Bias in the Validation of
ETS Language Tests
Jeffrey L. Zorn

I speak today as a college professor with a long-standing interest in fairer and more generally purposeful articulation between secondary and higher education, a problem whose scope has expanded geometrically with the opening of higher educational opportunity. I aspire neither to advance empirical knowledge about culture bias nor to rake anyone over the ideological coals. Rather, my aim will be Dristan's, to open blocked passages, in this case intellectually blocked passages to the investigation of culture bias in the Scholastic Aptitude Test, Verbal section; the English Composition Test; and the Test of Standard Written English.

A full examination of possible culture bias in these tests would cover many more areas of the psychometry than time here permits. That larger study would consider much about test preparation (both broadly and narrowly construed), the test items, the testing format, and the guidelines for evaluation and norming. In choosing to limit my comments to validation alone, I mean to encourage further detailed work in these four areas, not to imply their barrenness of interest a priori.

Predictive validity, though, is central to the significance of these and all other entrance tests. The SAT and the English Composition Test are tools of admissions work, the Test of Standard Written English a tool of English department administration; all three tests are employed as indices, predictors of achievements held to be of value in and of themselves, "the real thing," as one recent ETS informational booklet aptly

put it. ETS indeed provides a service, a psychometric service that applicants and administrators buy into in the confidence that the tests work. Comparing test score and criterion score is an essential task, for it is simply nonsense to claim, "We have a peachy-keen test here, but it doesn't measure, even indirectly, anything you're interested in."

The charge "culture bias in validation" turns, then, on these three matchups: the SAT-Verbal with college grades, almost always freshman grades; the English Composition Test with scores on essay-writing samples by college applicants; and the Test of Standard Written English with college-level writing performance measured by essay-writing tasks and English grades.

Researchers on culture bias in the SAT have looked for a consistent misprediction of grades made at college by students of different cultural background. The SAT-Verbal would be biased in its prediction function if, say, scores on it correlated much more strongly with freshman grades for white students than with grades for Black or Chicano students. Potthoff (1966) set the psychometric definition of "culture bias" in these terms: "A test is not biased if individuals from different groups who have the same test score also have the expected criterion score." In this view, of course, lower scores by a group constitute no prima facie claim of culture bias against the test, for the college grades made by these applicants, singly and collectively, might be lower in exactly the same proportion.

Many studies have been conducted to determine whether the SAT-Verbal predicts grades in the "college biased" fashion just described, and their general conclusion has been that it does not. Stanley and Porter (1967), Cleary (1968), Pfeiffer and Sedlacek (1971), and Medley and Quirk (1974) all detected no significant difference in the Verbal's prediction of black and white collegians' GPA's. Wilson (1969), Kallingal (1971),

Temp (1971), and Stanley (1971) all found evidence of a slight overpredic-tion of grades by black undergraduates. Goldman and Richards (1974) found similar overprediction of Mexican-American as against Anglo-American GPA's at UC-Riverside. Surveying evidence like this, Arthur Jensen in his most recent book Straight Talk about Mental Tests (1982) could crow, "Thus, contrary to popular claims, the SAT, in every case where it is a biased predictor, is biased in favor of black and Mexican-American students" (p. 148).

Arthur Jensen's words should not be the last on the subject of culture bias on the SAT. The discussion can be continued most profitably, I believe, by noting that the psychometric definition of "culture bias" itself is inadequate. It has blocked investigation, first, by excluding from vision all rejected applicants and, second, by accepting too readily the full value of college grades, especially freshman grades, as a significant and unbiased criterion of achievement.

The Scholastic Aptitude Test has but one clear purpose, to help admissions officers limit their mistakes, this by correcting for misleading high school transcripts. False-positive mistakes occur when a freshman is admitted and then does very poorly, maybe flunking out by the end of the freshman year. If the examination screens out some number of applicants who otherwise would have been admitted and done poorly at State, at Tech, and so on, it's done what it's supposed to. However, there is another kind of admissions mistake as well, false-negative selections, which occur when a young man or woman is deemed inadmissible who would in fact have succeeded at State, at Tech. These mistakes have received almost none of the attention that false-positive ones have gotten, probably because applicants once rejected are no longer the concern of the admissions professionals involved.

Using a host of innovative methods, some obviously tentative, Goldman and Widawski (1976) detected a consistent pattern of false-negative selection when the SAT was used in conjunction with high school transcripts at UCLA, UC-Davis, UC-San Diego, and UC-Irvine. Their strategy was to see which Black and Chicano applicants to these schools were admitted and how well they did, then to determine which of the rejected Black and Chicano students would have been admitted if high school record alone had been used as a screening measure, and then to predict success or failure for these students if they had been admitted, clearly the most controversial part of the project. Goldman and Widawski concluded that the use of the SAT yielded a reduction in false-positive selections for Blacks and Chicanos, but a far more pronounced increase in false-negative selections. They wrote: "The SAT produces a considerable reduction in the admissibility of would-be successful students. This occurs by the substitution of false-negative errors for false-positive ones."

Goldman and Widawski's innovative work should be developed much further, based as it is on the recognition of an entirely ignored party in culture bias research. Perhaps it will be found that the PSAT and SAT discourage many culturally different students even from applying to schools at which they would have succeeded, if not with great distinction, at least with lessons learned and a diploma in hand.

As SAT-Verbal scores have been declining precipitously the past 20 years, college grades have been inflated nearly half a grade on the standard four point scale. Surprisingly, the SAT has been shown to have supplemented the high school transcript increasingly well in this period. Bejar and Blew (1981) suggest cogently enough that grade inflation and curriculum

changes at the high school level have greatly reduced the statistical trustworthiness of transcripts, making the SAT's contribution higher by a kind of default. Still, the fickleness of college grading, over time, between schools, between divisions at the same school, even between sections of the same class, should make one skeptical about inferring one clear intellectual advantage held by a student identified only as having a higher GPA than a second student.

The College Board-commissioned panel investigating the SAT score decline of the sixties and seventies noted the grades up, tests down puzzle and provided strikingly good analysis:

The correlation between grades and scores and more grades could increase or decline or stay level, even while there was dangerous deterioration or healthy improvement by broader measures on both sides of the secondary/postsecondary divide. That correlation might come to apply to a smaller and smaller part of the learning actually going on. ("On Further Examination," P. 11).

The mention of "broader measures" brings to mind the language test on the National Assessment of Educational Progress. Both the reading and especially the writing sections of the National Assessment have been designed to let test-takers show what they can do with language, as opposed to testing and grading techniques that rank-order students by how much they don't and can't do. Precisely for this reason, one feels, National Assessment testing has shown the culturally different student, especially the not-best culturally different student, in better light than either school grades or the SAT.

In a recent American Scholar article, Jay Amberg argued that the SAT and the institutions of higher learning in this country are and should be culturally biased, biased, that is, in favor of higher literacy, rationality, and scholarship. Amberg hastily writes off the possibility of a con-

sistency misapplication of these glitteringly general concepts. Test writers and college teachers alike may well be rewarding mostly precociousness, trivial and evanescent advantages held by students bred to be "college material."

College-level English testing, at least, offers no shortage of instructive evidence of culture biased assumptions and not of the benign sort Amberg and I would laud equally. In concluding his detailed 1963 study of English instruction at the college level, for example, Albert Kitzhaber printed as an appendix two failing student essays, one from Dartmouth College, the other from an unnamed state university. Both are terrible papers, but for different reasons, the Dartmouth paper for its insincerity and lack of anything to say, the other for what we would call its illiteracies. Fully aware of how putrid the Dartmouth paper was, Kitzhaber was horrified by the second paper and deemed the teaching situation vis à vis its author "hopeless." He then indicated that higher education would do well to exclude all students who would commit such painful errors in spelling, punctuation, and word usage for their professors to read.

The good sense of these priorities is not immediately obvious. Perhaps the Dartmouth boy truly was the "hopeless" one, with the potential to supersede his badly expressed B.S. only with a smoother line of B.S. Perhaps the other student would have made excellent progress as a writer under gifted tutelage and eventually would have passed the Dartmouth boy by Dartmouth standards for reasons of greater intellectual character. Until quite recently college English teaching considered instruction in the mechanics of writing beneath it, but not instruction in elementary rhetoric, in how to sound not too callow, not too wooden. This ordering should be seen as a professional quirk, a prejudice, not the embodiment of serious cognitive task analysis: Compositions with

mechanical errors are not devoid of pedagogical interest; some, in fact, may have great literary interest, philosophical interest, and so on.

Insofar as the SAT is used to predict culture biased grades, it dis-serves certain college applicants, certain of its clientele, regardless of the cogency of its own internal controls. Testing professionals should want to probe the culture-fairness of the SAT's criterion for this reason.

Almost exactly contemporaneous with Kitzhaber's study was the College Board's Research Monograph #6, "The Measurement of Writing Ability." whose methodologies and findings underpin the other two language tests under consideration. Both these tests set multiple choice questions on usage, number agreement, and so on; and results on both are understood as predictors or indices of actual writing ability as seen on essay tasks. Research Monograph #6 provided the rationale for this view of objective testing in relation to every testing.

In-house ETS researchers Godshalk, Swineford, and Coffman set out to validate the 20-minute objective tests that make up the English Composition Test against a refined criterion of writing ability. Their plan was to test 646 high school juniors and seniors one way (objective tests) and then another way (short essays) and compare the results. Godshalk et al. set the firmest-ever statistical base for the validation project by using more essays for each subject and more readings by trained evaluators of each essay. The correlations they obtained were strong enough for them to conclude that colleges confidently could take scores on the English Composition Test as accurate measures of their applicants' relative ability in writing.

Of primary interest to the investigation of culture bias would be the 3-point holistic scale that the researchers trained the evaluators to employ.

In the words of Godshalk, Swineford, and Coffman: "The readers were asked to make global or holistic, not analytical judgments of each paper, reading rapidly for a total impression. There were only three ratings: a score of "3" for a superior paper, "2" for an average paper, and "1" for an inferior paper." Those concerned for the fair treatment of culturally different students would do well to press researchers and English educators precisely at this general point: What is to count as superior, average, and inferior performance? Does Black English, e.g., ensure its user a "1"? Can cheerily voluble standard English get anything but a "3"?

Godshalk, Swineford, and Coffman set an unfortunate methodological precedent in proceeding as if the move to holistic scoring had relieved the study of the burden of setting standards, as if evaluation with overt reference to wholes and not parts somehow is bereft of judgment, a-criterial. Globally I can say of a play, a symphony, a basketball game, "It stunk," but if pressed I should be prepared to detail the missed cues, missed notes, and missed free throws that informed the larger judgment. I may have forgotten some of these key details, but I appealed to them, and only to them, as I was preparing the global statement. Similarly, holistic evaluation of essays builds on a series of analytical judgments, however fleetingly considered and however fully ignored in the public reporting of reaction. These atomistic judgments and informal weighing formulae can be stated rather more explicitly than we are accustomed to doing and then debated; if need be, they can be reformed.

Also, the 646 students in the Godshalk study represented no true cross-section of college applicants. Only 24 secondary schools were tapped, with classes of average ability and achievement selected. Curiously, Godshalk, Swineford, and Coffman understood this very restricted sampling as able

"to avoid the possibility of a limited range of testing results." The authors ingenuously described the student sample as representing "all commonly listed geographical areas of the country," somehow thinking of geography as the key to range. Setting another poor example for method in ETS validation study, Research Monograph #6 dodged the question of standards for the culturally different student simply by not considering that student.

A similar criticism can be made of Hunter Breland's replication of the Godshalk study, in-house work that directly undergirds the use of the Test of Standard Written English as part of the Scholastic Aptitude Test. Among test-takers with a lifetime of exposure to the language conventions under examination, the test stands a good chance of picking out stronger and weaker writers. Culturally different students have no such linguistic background, however, and so the test stands in qualitatively different relationship to their writing than to the writing of the "prepped" students. Too easily the Test of Standard Written English could lead to the classification of all culturally different students as "remedial" students, regardless of their true abilities and deficiencies as composers. Again, the difference between "remedial" and "expected and therefore non-remedial" shortcoming is everywhere the product of evaluative theories; these mustn't be smuggled in as somehow following from certain charts of correlation.

English teaching, and test-writing supporting English teaching, must have greater sociological imagination than this. The profession does not yet know as much as it should about the developmental paths to excellence taken by culturally different youth. As we strive to learn more about these paths and the roles we teachers can play to accelerate students down them, we must also work to brand as invidious the legislation of a single path to higher literacy, especially one marked by the traits of those young people with whom the legislators are most familiar and most comfortable.